Auditory-Visual Integration of Sine-Wave Speech

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for Graduation with Distinction in Speech and Hearing Science in the Undergraduate Colleges of The Ohio State University

By

Matthew Joseph Tamosiunas

The Ohio State University

June 2007

Project Advisor: Dr. Janet M Weisenberger, Department of Speech and Hearing Science

Abstract

It has long been known that observers use visual information from a talker's face to supplement auditory input to understand speech in situations where the auditory signal is compromised in some way, such as in a noisy environment. However, researchers have demonstrated that even when the auditory signal is perfect, a paired visual stimulus will give rise to a different percept from that without the visual stimulus. This was demonstrated by McGurk and McDonald (1976) when they discovered that when a person is presented with an auditory CV combination (e.g., /ba/), and visual speech stimulus (e.g., /ga/), the resulting perception is often a fusion (e.g., /da/) of the two. This phenomenon can be observed in both degraded and non-degraded speech stimuli, suggesting that the integration is not a function of having a poor auditory stimulus.

However, other studies have shown that the normal acoustic speech stimulus is highly redundant in the sense that the signal contains more information than necessary for sound identification. This redundancy may play an important role in auditory-visual integration.

Shannon et al. (1995) reduced the spectral information in speech to one, two, three, and four bands of modulated noise using the original speech envelope to modulate the same spectral band. The results showed very high intelligibility even for reductions to three or four bands, suggesting that there are tremendous amounts of redundancy in the normal speech signal. Furthermore, Remez et al. (1981) reduced the speech signal to three time-varying sinusoids that matched the center frequencies and amplitudes at the first three formants of the natural speech signal. Again, the results showed high

intelligibility (when the subjects were told that the sounds were, in fact, reduced human speech).

A remaining question is whether reducing the redundancy in the auditory signal changes the auditory-visual integration process in either quantitative or qualitative ways.

The present study addressed this issue by using, like Remez, sine wave reductions of the auditory stimuli, with the addition of visual stimuli. A total of 10 normal-hearing adult listeners were asked to identify speech syllables produced by five talkers, in which the auditory portions of the signals were degraded using sine wave reduction. Participants were tested with four different sinewave reductions: F0, F1, F2, and F0+F1+F2. Stimuli were presented under auditory only, visual only, and auditory plus visual conditions.

Preliminary analysis of the results showed very low levels of performance under auditory only presentation conditions for all of the sinewave reductions, even F0+F1+F2. Visual-only performance was approximately 30%, consistent with previous studies. Little evidence of improvement in the auditory plus visual condition was observed, suggesting that this level of reduction in the auditory stimulus removes so much auditory information that listeners are unable to use the stimulus to achieve any meaningful audiovisual speech integration. These results have implications for the design of processors for assistive devices such as cochlear implants.

Acknowledgments

- I would like to thank Dr. Janet Weisenberger, for being my thesis advisor and for helping me to achieve such a wonderful academic experience and inspiring me to pursue future research in audiology.
- I would like to thank Natalie Feleppelle for the tremendous amount of assistance during laboratory preparation, testing, data analysis, poster preparation, and presentation; all with an amazing amount of enthusiasm.
- Finally, I would like to say thank you to The Ohio State University and the College of Social and Behavioral Sciences and Department of Speech and Hearing Science for making such a great opportunity possible for undergraduate students to better prepare us for the future.
- This project was supported by an ASC Undergraduate Scholarship and an SBS Undergraduate Research Scholarship.

Table of Contents

Chapter 1: Introduction and Literature Review

We generally believe speech perception to be a phenomenon that begins with the ears collecting sound and the brain translating it into language. This holds true in most cases, but when do the eyes play a role in speech perception? Our ears are naturally tuned to the frequencies produced by human speech, and so it makes sense that they would be considered a primary receiver for speech, but in most environments, ears alone are not enough help due to noise and/or hearing loss. Our eyes compensate for this loss of sound and receive articulatory information produced by the mouth during speech. These two mechanisms work together in harmony to allow us to receive information from other humans in acoustically difficult environments. The resulting perception (human speech) is gained by this auditory-visual integration.

It has long been known that auditory-visual integration occurs when the auditory signal has been compromised in some way, as in a noisy environment. However, we now know that this integration occurs even when the auditory signal is perfect due to the work of McGurk and MacDonald (1976). In their study, certain auditory information was overlaid onto non-matching visual information, creating a discrepant speech signal. For example, an auditory /ba/ was dubbed onto a visual /ga/. The resulting perception happened to be /da/, a fusion of the place of articulation of the two syllables. The syllable /ba/ is considered to be bilabial (articulation of both lips) while /ga/ is velar (articulation of the velum). The resulting /da/ is alveolar (articulated at the alveolar ridge) which falls in between the other two places of articulation. This fusion of auditory

and visual inputs occurs even when the auditory information has not been compromised in any way, indicating that visual data is being used in the perception process and that the observer cannot ignore this visual input.

Visual Cues for Speech Perception

The knowledge of place of articulation can be obtained from movement of the talker's eyes, mouth, and head (Munhall et al., 2004). In many circumstances the visual representations of sounds have similar visual characteristics. A viseme is defined as a basic unit of speech in the visual domain, while a phoneme is a basic unit of speech in the auditory domain. A viseme group generally corresponds to at least one phoneme, and usually more. For example, the phonemes, /p, b, m/ are often considered one viseme because they are all stops produced in a bilabial manner and cannot be distinguished by sight alone; auditory information is necessary to distinguish any of these from the other. Without any auditory-specific information (manner and voicing), an observer would have a difficult time distinguishing among them. /d, t, n/ and /k, g/ are other common visemes. It is important to note that visemes are not universal, because of talker differences in articulation (e.g., one talker may show extreme 'plosivity' in their production of /p/, distinguishing it from /b/ and /m/, while many other talkers may not). Vowels can be grouped into viseme categories as well, but differences across talkers when producing those vowels cause confusion more often and, as a result, universality is harder to achieve than it is with consonants. Hard-to-speechread talkers will usually provide a smaller amount of viseme groups (Jackson, 1988) than will highly intelligible talkers.

Auditory Cues for Speech Perception

The auditory component of speech conveys place, manner, and voicing information to the listener through spectral and temporal aspects of the speech waveform. All of this information contained in the speech waveform is accompanied by much more information which, many researchers suggest, is redundant (more information is present than necessary for correct sound identification.) This suggestion stems from a multitude of experiments in which speech signals were degraded to various degrees and high intelligibility was still achieved. One such example was demonstrated by Shannon *et al*. (1998) when the spectral information in speech was reduced to two broad noise bands and then modulated by the original envelope. Results showed that recognition of vowels and consonants was greater than anticipated. Reduction of spectral information to four noise bands resulted in even greater recognition of the acoustic information, suggesting that almost all of the manner and voicing information was being conveyed. Shannon also concluded that recognition of consonants is less affected than recognition of vowels when such degrading techniques are used. Ultimately, this work demonstrated the possible redundancy and robustness in speech and would lead to other techniques of signal degradation.

Remez *et al.* (1981) reduced speech signals (utterances) to three time varying sinusoids centered on the first three formants (f0, f1, and f2). Sentences were presented to test subjects as a combination of all three sine waves which were perceived as three separate tones. Some subjects were told beforehand that the stimulus would be a speech utterance while others were not given any information. The results showed that, despite the tremendous lack of information, those with knowledge of the stimulus had very

accurately described the utterances while those with no information still detected some linguistic content.

Bistability of Sinewave Speech

Remez *et al.* (2001) studied how perceptual organization differs between synthetic speech and sinewave speech. In the first part of the experiment, subjects were asked if the two isolated second formant tones being presented were the same or different. They were then presented with a second formant tone followed by a word and asked if the tone was a part of the word. Scores were high in the first task, but very low in the second task, suggesting that auditory and phonetic organization were dependent upon one another. Julesz and Hirsh (1972) explained this idea of interdependence of auditory and phonetic perceptions as "perceptual coherence."

The second part of the study examined subject performance under the sinewave condition F2. First the subjects were presented with a sample F2 tone followed by a (sinewave) tone complex possibly containing the sample tone, and asked whether the tone was a part of the complex. The subjects were then presented with a printed word, and sample F2 tone followed by a (sinewave) tone complex possibly representing the written word (and sample F2 tone). Scores were high for all tasks and lexical verification scores in the final part showed very high scores. These results suggest that auditory and phonetic organization occur simultaneously in sinewave speech.

Overall, findings showed that, unlike synthetic speech (in which auditory and phonetic organization cohere), sinewave speech is "perceptually bistable," meaning that "phonetic organization of sinewave analogues occurs independently of auditory

organization" (Remez, 2001, p. 29).

Measures of Auditory-Visual Integration for Hearing Impaired Listeners

It is known that speech is perceived more accurately when both an audio and visual stimulus are presented together, as opposed to audio-alone or visual-alone states. This benefit achieved through auditory-visual integration was studied by Grant and Seitz in 1998 with hearing impaired individuals. By presenting nonsense syllables to each participant in A, V, and A+V conditions, a measurement of AV benefit was taken after comparing A to A+V and V to A+V. Results showed that even though hearing was impaired, listeners displayed significantly high AV benefit.

Auditory-Visual Integration Theories

Two theories that were developed to determine the ability to optimally integrate the auditory and visual systems are worth discussing. The pre-Labeling Model of Integration (PRE) was developed by Louis Braida, and predicts how well a person should be integrating both modalities after collecting data on visual alone and auditory alone capabilities (Cited in Grant, 2002). Theoretically, the auditory-visual (AV) scores should be equal to or exceed the recognition scores for auditory-alone (A) and visual-alone (V). If the AV scores fall within prediction of the model then the individual is said to be integrating efficiently and rehabilitation ought to be focused on A and/or V recognition alone. If the AV scores fall below the predicted scores then the individual is said to not be integrating both modalities efficiently and, therefore, rehabilitation should be focused on integration training. Grant also notes that hearing impaired listeners are generally

over-predicted using this model.

The Fuzzy Logical Model of Perception (FLMP) was developed by Massaro to "fine tune" the PRE model by reducing the variation between the predicted and obtained recognition scores (Grant, 2002). Grant, however, disagrees with the FLMP because, in contrast to the PRE, it underestimates integration abilities.

Auditory-Visual Integration Efficiency in Normal and Hearing Impaired Listeners

In 2007, Grant published findings from a study comparing the auditory-visual integration benefit of normal hearing individuals and hearing impaired individuals. As in his 1998 study, he presented listeners with nonsense syllables in A, V, and A+V conditions. Audio stimuli were reduced using four nonoverlapping filter bands between 300 and 6000 Hz. Both groups displayed significantly high AV benefit, but the difference between the groups was that hearing impaired individuals displayed less integration across the auditory-only condition (or across the acoustic frequency spectrum).

Role of Redundancy in Auditory-Visual Speech Perception

Auditory speech signal redundancy was demonstrated by Shannon et al. (1995) when he and his colleagues reduced the spectral information of speech while manipulating the temporal envelopes to preserve temporal cues. As predicted, greater speech recognition resulted when a greater number of noise bands were used, but high recognition resulted with as little as three bands. Surplus acoustic information, therefore, is believed to be present.

In 1981, when Remez et al. degraded the speech signal (utterance) to three sine waves centered on the first three formants, individuals still recognized linguistic content despite sine wave reduction being among the most impoverished auditory signals. Once more, evidence of surplus information was presented.

We also know from the studies of Jackson that while one is speechreading, it is hard to distinguish between phonemes (due to place of articulation being the only cue), resulting in perception of a viseme group at best. This results in ambiguity of the visual speech signal.

To understand the role of redundancy in auditory-visual speech perception it is important to determine how the degree of redundancy in the auditory signal affects the strength of the McGurk effect. Although both the auditory and visual speech signals convey information on place of articulation, there is redundancy in the speech signal and ambiguity in the visual signal. The unanswered question is: What circumstances promote optimal auditory-visual integration? Is a certain degree of redundancy necessary for integration to occur and, if so, how much? The answers to these questions may be answered by stripping varying amounts of redundancy from the auditory speech signal and observing the degree of resulting integration of both modalities.

Previous research has demonstrated three important facts: 1.) auditory-visual integration is extremely beneficial when the auditory signal is compromised in some way, 2.) human speech is redundant (acoustically) as well as ambiguous (visually), and 3.) The McGurk effect shows that visual input is used even when auditory input is perfect when perceiving speech.

The present study investigated how auditory-visual integration occurs for isolated

CVC syllables by presenting highly reduced, non-redundant speech cues in the form of sine waves together with visual speech information. Ten normal hearing adults were asked to identify speech stimuli under three conditions: auditory alone (A), visual alone (V), and auditory + visual (AV). Under the AV conditions, both congruent (Matching A and V phonemes) and discrepant (A phoneme is different from V phoneme) combinations were presented. Results of this and future studies should have implications for signal processing strategies for hearing aids and cochlear implants as well as for designs for rehabilitation programs.

Chapter 2: Method

Participants

Ten adult students (8male, 2 female) between the ages of 20 and 24 participated in this study, all of whom reported normal hearing and vision. Five of the ten participants had taken an introductory phonetics course while none of the others had any linguistic background. In addition, five more participants (2 male, 3 female) between the ages of 19 and 24 were video-recorded to provide the stimuli being presented. Each participant received $80.00 for their time.

Interfaces for Stimulus Presentation

Visual Signal Presentation

Each participant sat in a chair inside a sound attenuated chamber. A 50 cm video monitor was placed 60 cm outside the window of the chamber at eye level which was about 4 feet from the participant's face.

Degraded Auditory Signal Presentation

The physical setup for degraded audio signal presentation remained the same as for the visual signal presentation except that instead of observing the video monitor, the audio signal was sent to the participant via 600-ohm TDH circum-aural headphones. The monitor was turned off and a shade was pulled down to prevent visual distraction.

Visual + Degraded Auditory Signal Presentation

In this condition, the monitor was visible and the headphones were worn by the participant to allow the use of both modalities.

Stimuli

Stimulus Selection

A set of eight CVC syllables were used as the stimulus for this study. Each syllable was selected in accordance with the following conditions:

1.) Pairs of the stimuli were minimal pairs, differing by only one phoneme: the initial consonant.

2.) All stimuli were accompanied by the vowel /æ/, since it does not involve lip rounding or lip extension.

3.) Multiple stimuli were used in each category of articulation, including: place (bilabial, alveolar), manner (stop, fricative, nasal), and voicing (voiced, unvoiced).

4.) All stimuli were presented without the use of carrier phrases (citation style).

5.) The stimuli were known to elicit McGurk type responses.

Stimuli

Random orders of the same 8 stimuli were used in each condition. These include:

1.) Bat

2.) Cat

3.) Gat

4.) Mat

5.) Pat

6.) Sat

7.) Tat

8.) Zat

Stimulus Presentation

Audio Signal Degrading

Five talkers provided the speech stimuli used in this study by talking directly into a microphone connected to a computer allowing *Video Explosion Deluxe* software to save each recording as a .wav extension sound file. Each talker repeated the selected set of (eight) CVC syllable stimuli five times. These audio files were then degraded to three sine waves centered on the first three formants (F0, F1, and F2) using *Praat version 4.4.29* software and also using a script developed by Chris Darwin of The University of Sussex. The program reads the specified audio file (e.g., .wav) and converts it to sine waves based on the gender and age of the talker. The upper formant limits used were 5000 Hz for an adult male and 5500 Hz for an adult female.

Digital Video Editing

Visual stimuli were obtained by digitally video-recording five talkers (2 male, 3 female) repeating the list of eight CVC stimuli five times each. The stimuli were then transferred to the hard drive and were thus accessible to *Video Explosion Deluxe* software.

With both audio and visual stimuli accessible to *Video Explosion Deluxe,* the video files were created. A video file (.avi) of a talker was selected and a corresponding audio file (.wav) of the same talker (for the purposes of this study) was dubbed onto it. Some auditory-visual stimuli were congruent (matching A and V phonemes), while others were discrepant (non-matching A and V phonemes). The discrepant condition would allow for elicitation of McGurk responses.

Once the movies were compiled and created into .avi format, *NeroVision Express*

3 software was used to burn them onto DVDs. The set of DVDs for the experiment consisted of sixty DVDs. Each of the five talkers was used in three DVDs, to allow for an A, V, and A+V component (DVD) for each of the four sine wave configurations. The DVD used in each trial was randomly selected to reduce learning effects.

The testing was divided into three presentation conditions for each sine wave configuration, which included visual only, sine wave (degraded) audio only, and sine wave (degraded) audio plus visual. The order of conditions was randomized across all participants. Each trial consisted of the participant repeating the syllable that he/she believed was being presented while the examiner recorded the responses.

Procedure

Testing Setup

Testing for the present study took place in the Audio-Visual Integration Research Laboratory of the Department of Speech and Hearing Science at The Ohio State University. The room provided a quiet and well lit atmosphere conducive to research of the present type. The participants sat in a chair in a sound attenuated chamber facing a video monitor for visual presentation through the window of the chamber. Headphones were wired through the chamber for audio presentation. Communication between participant and examiner took place via intercom system installed on both sides of the chamber.

Once testing was initiated, the chamber door was sealed and the shade of the chamber window was lowered or raised accordingly. When the condition being presented was video only, the headphones were unplugged so as not to provide an acoustic signal. When the condition was audio only, the headphones were plugged back

in, the shade was pulled down, and the video monitor was turned off.

Testing Tasks

Each DVD presented the participants with 60 randomly ordered stimuli consisting of the aforementioned CVC syllables. This DVD was also randomly assigned to be used for only one condition (A, V, or A+V) for each participant. After the presentation of each of the 60 syllables the participant was given the opportunity to tell the examiner what syllable they believed was being presented based on the condition at hand (A, V, or A+V). The examiner recorded each response on paper data sheets corresponding to each DVD. These data were later transferred to *Microsoft Excel* software for analysis.

Testing Presentation:

Testing consisted of three conditions utilizing the 60 prerecorded DVDs. The audio alone and visual alone conditions were composed of 60 congruent stimuli on each DVD while the audio + visual condition used only 30 congruent stimuli. The remaining 30 were discrepant stimuli used for purposes of eliciting McGurk responses. Learning bias and stimulus memorization problems were eliminated through the use of randomization of DVDs for all conditions and all participants.

Testing Procedure:

Each participant was tested under all three stimulus presentation conditions. The presentation conditions were audio only, visual only, and audio + visual. Each condition consisted of five talkers presenting stimuli in each of the four sine wave configurations, totalling 60 DVDs. The sine wave configurations consisted of F0, F1, F2, and F0+F1+F2. The stimuli were recorded onto DVD and presented via the video monitor. Each test session lasted 2 hours with a rest period every half hour.

Chapter 3: Results and Discussion

Results were analyzed for two types of stimuli: congruent-syllable presentations (auditory syllable was paired with the same visual syllable), and discrepant-syllable presentations (auditory syllable was paired with a different visual syllable). In congruent-syllable presentations, degraded auditory-only, visual-only, and degraded auditory + visual conditions were each assessed for performance in all four sinewave conditions (F0, F1, F2, and F0+F1+F2) which was done by calculating the percent of correct responses in each situation. Auditory-visual integration can be measured by comparing performance between degraded auditory-only and degraded auditory + visual conditions (Figure 1) or by comparing between visual-only and degraded auditory + visual conditions (Figure 2). Degraded auditory + visual conditions are assumed to produce higher percent correct scores thus reflecting integration of modalities.

In discrepant-syllable presentations (e.g., auditory /ga/ paired with visual /ba/) responses were categorized into three categories: auditory, visual, or other (Figure 3). Responses falling into the category 'other' were further analyzed to determine if they fell into one of two sub-categories: fusion (of the places of articulation) or combination (of the places of articulation). In the case of the discrepant-syllable example above, a valid fusion of auditory /ba/ and visual /ga/ would be /da/ and a valid combination (addition of the places of articulation) would be /bga/. Fusions and combinations are considered to be the results of auditory-visual integration. This study did not observe any combination responses. Figure 4 depicts the percentage of "fusion" responses to "neither" (neither

fusion nor combination) responses.

Percent Correct Identification from Congruent Stimuli

Analyses were done using 2-factor within subjects ANOVA (arcsin transformed data). Figure 4 shows the percent correct responses by sinewave configuration under auditory-only, visual-only, and auditory + visual conditions. The particular sinewave configuration did not seem to affect performance; no significant main effect of sinewave condition was found, $F(3,147) = .57$, ns, and scores by presentation condition seemed to be consistent across sinewave configurations. Visual-only scores (consistent with previous studies at approximately 30%) were far higher than auditory-only scores, reflecting the high acoustic data reduction. However, auditory + visual scores were lower than visual-only scores, suggesting that the auditory signal was degraded to such an extent that it may have interfered with normal visual perception rather than being integrated with it. ANOVA showed a significant main effect of presentation condition, $F(2,98) = 208.4$, $p<.001$, $\eta 2 = 81$. Followup pairwise comparisons showed significant differences across all modalities. Finally, a minimally significant interaction effect was observed, $F(6,294) = 2.3$, $p = .049$, $\eta 2 = .045$. However, this finding was likely not attributable to the sinewave manipulation.

Results also showed very low levels of performance in the auditory-only condition for all sinewave configurations. The surprisingly poor performance of listeners with these sinewave stimuli suggests that previous results of Remez et al. (1981) with sinewave sentences were dependent on the acoustic variation and linguistic content of the sentence stimuli. In the present study, 5 of the 8 CVC syllables are common English

words; these were most often identified correctly by listeners.

Figure 5 shows results for each talker. Although little variation is seen in auditory performance, differences are apparent in the auditory + visual condition. However, Figure 2 indicates that much of this variability is explained by visual-only performance. For Talkers 2 and 3, the addition of the auditory signal seemed to add particular interference.

Observers typically improve in sinewave speech perception after extensive exposure to it. Future work could investigate the impact of training on audiovisual integration of sinewave syllables. In addition, a stimulus set employing all words that varied in both initial consonant and medial vowel might yield higher levels of performance. Overall, the present study suggests that when too much information is removed from the acoustic stimulus, listeners are not able to use it in auditory-visual integration of speech.

Percent Response from Discrepant Stimuli (McGurk Stimuli)

The remaining analysis consisted of discrepant stimuli (in which the auditory syllables did not match the visual syllables). Figure 3 shows that about 36% of the responses were decided by the visual modality and only about 8% by the auditory modality. Furthermore, the remaining 'other' percentage of about 56% was subdivided into two more categories: fusion and neither (no combinations were found in this study). Figure 6 shows that compared to about 10% fusion responses, 'neither' responses dominated at around 49%. This may suggest that, due to the great amount of missing auditory information, the auditory component of the stimuli did not carry the necessary

information of integration. Also, the auditory stimulus may have stripped of so much information that, as mentioned before, the visual percept was affected by it.

When analyzing these results by talker, fusion seemed to vary (fig. 6). In fact, it varied chronologically; the first talker showed the most fusion responses, the second talker showed the second most fusion responses, and so on. This may suggest that listener attention played a role in integrating auditory and visual modalities during sinewave speech.

Overall, the sinewave reductions reduced the redundancy in auditory speech signals to the degree that it may have affected overall auditory-visual integration. Observations from the discrepant syllable tests show that integration by way of fusion was minimal and suggests that sinewave speech is too degraded a signal to facilitate auditory-visual integration.

Questioning Poor Identification Performance

As previous studies have shown low auditory performance in sinewave speech, this study indeed expected similar results. However, once the auditory + visual performance was observed to be poorer than visual-only performance (in every sinewave configuration) there was a need to look back at the methods for the study because it seemed surprising that the addition of a visual stimulus to an auditory stimulus would produce scores lower than the visual-only scores.

One possible aspect of the results may have been that the set of tokens did not vary in vowel; sinewaves in this study were produced at each of the first three formants which also happen to be the spectral regions that determine individual vowels.

Therefore, this auditory reduction method was attempting to reduce consonants in a way that is probably more effective for reducing vowels. In addition, vowels made up a great majority of each stimulus used in this study and so the acoustic information reduced was mostly vowels. Also, the addition of F3 to the stimuli would be beneficial in the future for the sake of certain obstruents (especially fricatives) that can not be identified at lower frequency regions.

Second, the selection of tokens used in this study included six words of Standard American English, which poses a possible obstacle when attempting to elicit auditory (not phonetic) responses in a study such as this one. Remez et al. (2001) provided evidence that listeners use both auditory and phonetic organization when perceiving sinewave speech. The use of these words may have inadvertently persuaded the listener to perceive more phonetically than auditorily, thus hindering the auditory-visual integration that was originally expected.

Chapter 4: Summary and Conclusions

Results of this study indicate that sinewave reduction of speech effectively reduces available acoustic information found in the signal. This is supported by the fact that only 13% of the auditory-only stimuli in the study were correctly identified by listeners. This study also suggests that there may not be enough information (redundancy) contained in sinewave speech to facilitate optimal (or any) auditory-visual integration because auditory + visual performance was lower than visual-only performance across all sinewave configurations. Finally, results of this study also support the idea of sinewave bistability.

Understanding how much information is lost from reduction to sinewave speech may be important in understanding how much redundancy is necessary in facilitating optimal auditory-visual integration. This knowledge may then be further utilized to improve aural rehabilitation programs.

Knowledge of how sinewave speech is perceived is a key component to understanding how auditory and phonetic organization works. A better understanding of sinewave perception may ultimately impact computer voice recognition systems. In fact, an automatic speech recognizer was built by Barker and Cooke (1997) that performs well in synthetic speech scenarios, but poorly in sinewave speech environments. However, training has proven to increase the recognizer's performance of sinewave speech.

Training human subjects in sinewave speech generally increases performance. The particular factors that comprise sinewave speech as being perceivable by humans

must be studied further so that those components may be compared to synthetic speech. Advances of this nature of study would have implications for cochlear implants and other assistive listening devices.

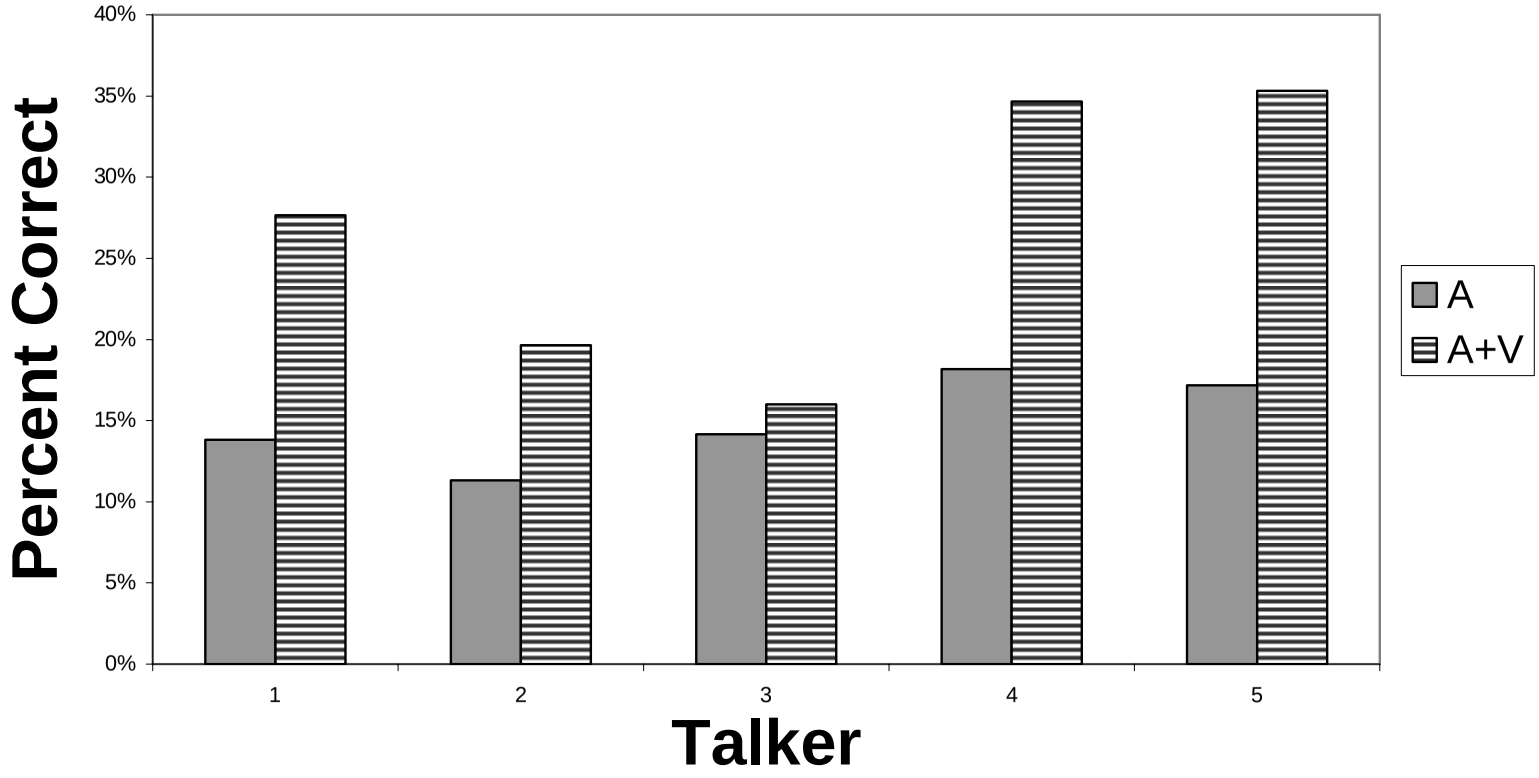
Chapter 5: References

- Grant, K.W. (2000). The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America, 109 (5)*, 2272 – 2275.
- Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *The Journal of the Acoustical Society of America, 112 (1*), 30 – 33.
- Grant, K.W. & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104 (4)*, 2438 – 2449.
- Grant, K.W., Tufts, J.B., & Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals. *The Journal of the Acoustical Society of America*, 121 (2), 1164 – 1176.
- Jackson, P.L. (1998). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90 (5)*, 99 – 114.
- Julesz, B., & Hirsh, I.J. (1972). Visual and auditory perception – An essay of comparison. In E.E. David & P.B. Denes (Eds.), *Human communication: A unified view* (pp. 283 – 340). New York: McGraw-Hill.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

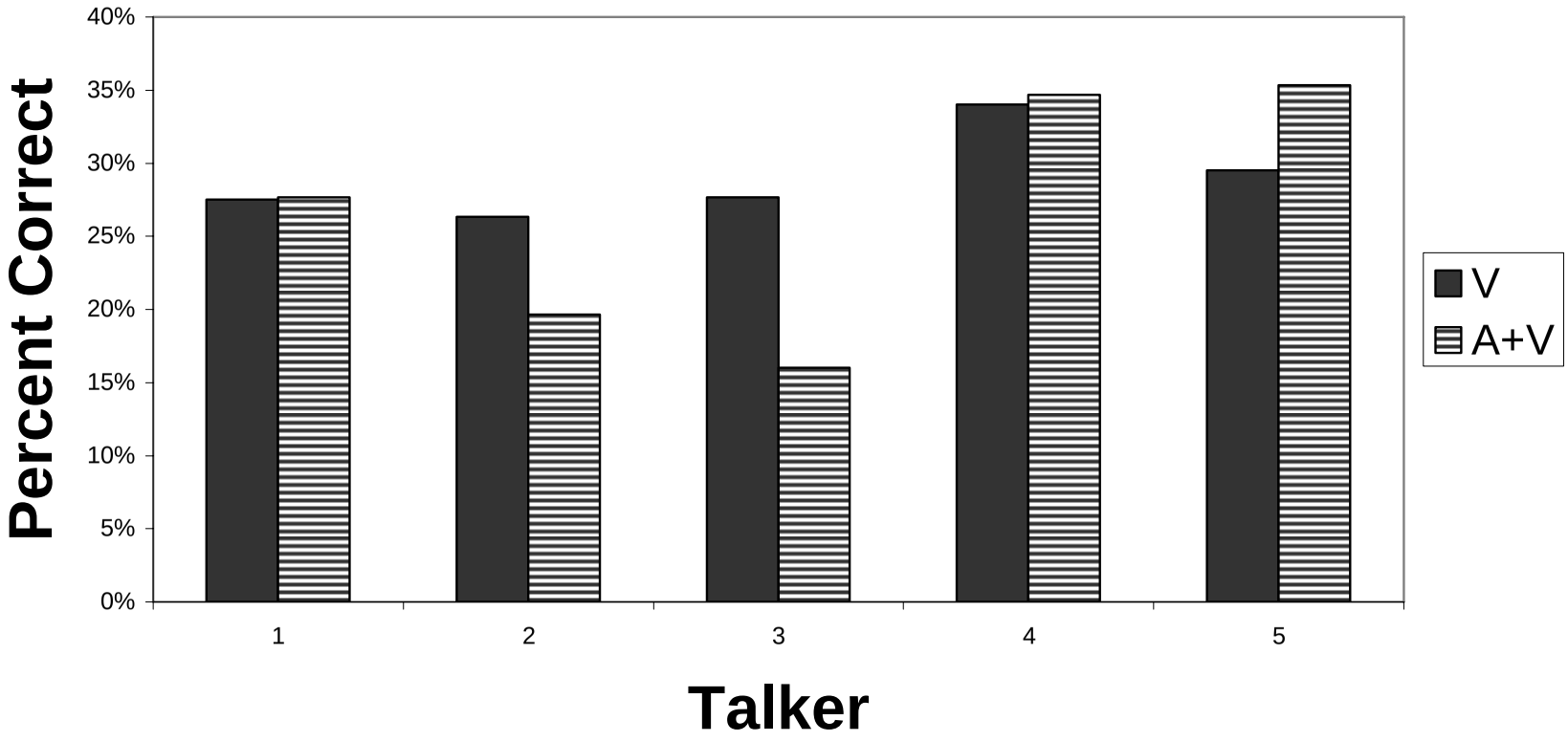
- Munhall, K.G., Kroos, C., Jozan, C., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perceptions & Psychophysics, 66 (4)*, 574 – 583.
- Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science, 212 (4497)*, 947 – 949.
- Remez, R.E., Pardo, J.S., Piorkowski, R.L., Rubin, P.E. (2001). On the bistability of sine wave analogues of speech. *Psychological Science, 12 (1)*, 24 – 29.
- Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303 – 304.
- Shannon, R.V., Zeng, F.G., Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *The Journal of the Acoustical Society of America, 104 (4), 2467 – 2475.*

List of Figures

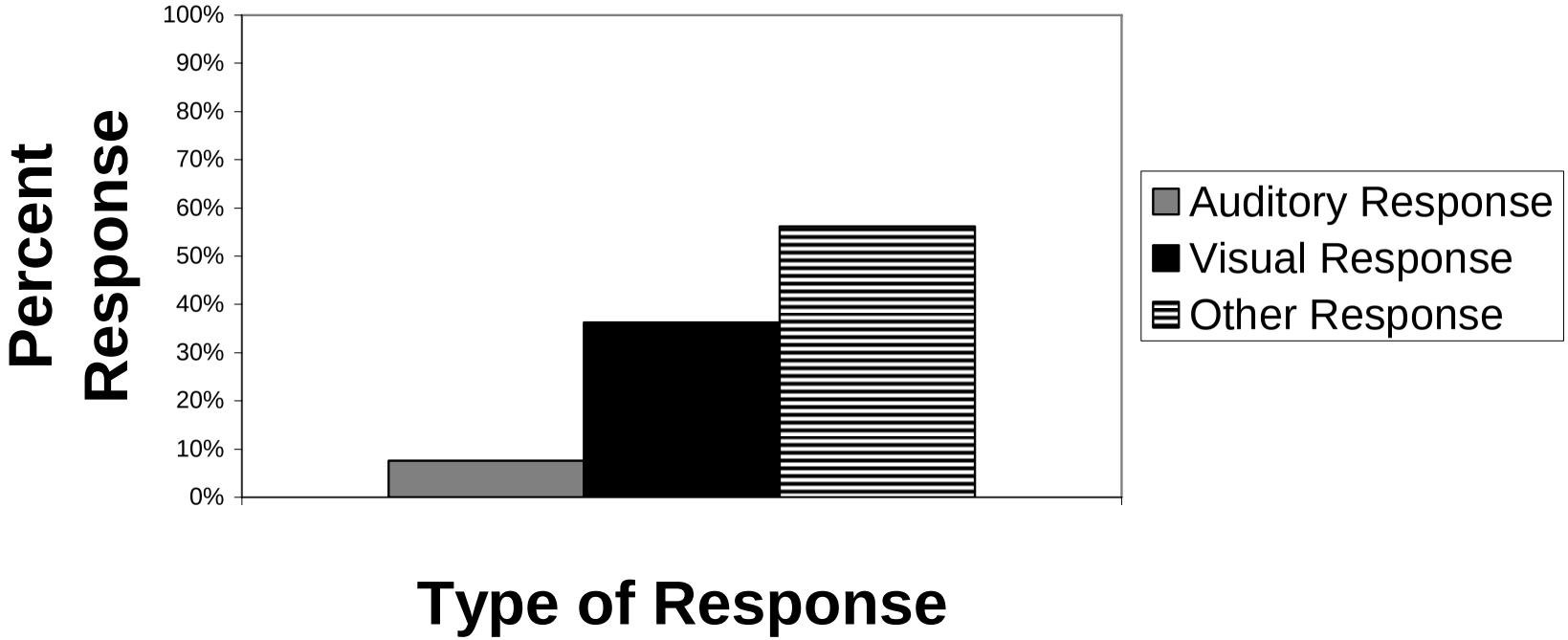
Percent Correct Auditory Only and A+V, F0+F1+F2



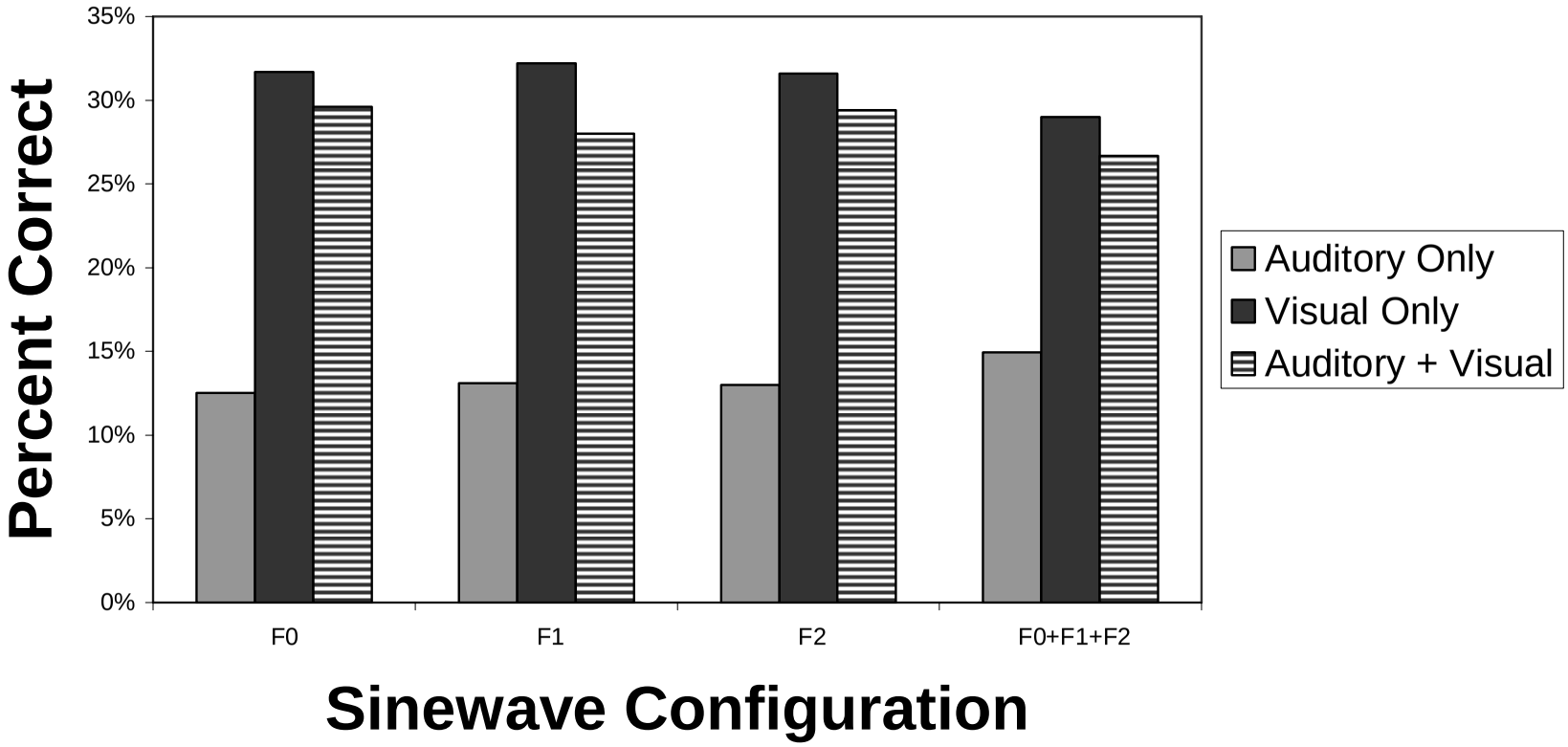
Percent Correct Visual Only and A+V, F0+F1+F2



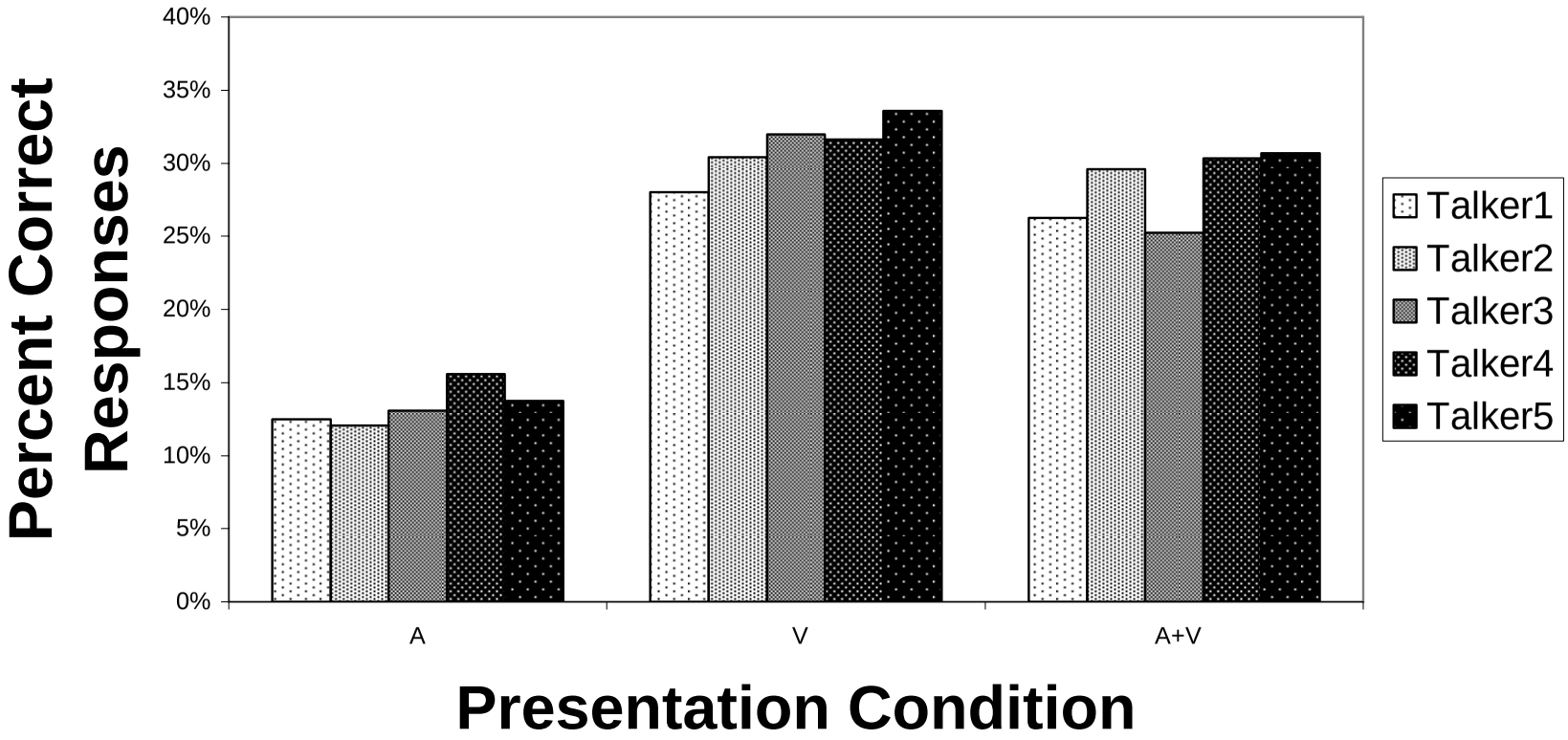
Percentage of Types of Responses From McGurk Stimuli



Overall Performance Across Sinewave Conditions



Performance of Talkers by Condition



Percent Fusion or Neither Responses by Sinewave Configuration